Genotyping of HIV-1

Mika O. Salminen, Jean K. Carr, Donald S. Burke, and Francine E. McCutchan

The Henry M. Jackson Foundation Research Laboratory and Division of Retrovirology, Walter Reed Army Institute of Research, Rockville Md. 20850

Introduction

The designation "Human Immunodeficiency Virus Type-1" (HIV-1) encompasses an unanticipated complexity of viral forms. The eight prevalent major (M) genotypes and a low prevalence, but perhaps equally diverse, outlier (O) group, have been described in this compendium [1]. Intergenotypic recombinants of group M viruses, often consisting of complex mosaics, are also recognized [2–5]. As this diversity unfolds, there is a continuing requirement to organize genetic data into a useful structure for the exploration of biologic and immunologic relationships among HIV-1 isolates. Here we describe some complications arising out of the genotypic substructure of HIV-1.

The majority of newly derived HIV-1 sequences are expected to belong to recognized group M subtypes or to group O. Genotypic assignments can be derived by parsimony, neighbor-joining, or maximum likelihood methods using available reference isolates. Occasionally, a sequence is found that does not cluster with recognized genotypes or that associates weakly with one of them, forming a branch nearer the main trunk. Figure 1 illustrates several isolates of this nature in published trees derived from complete gag or env gene sequences [6,7]; examples include isolates BZ200 and VI325 in the gag tree and isolates ZAM184 and VI525 in the env tree. Moreover, bootstrap resampling frequencies that are lower for some branches in the tree may indicate that mosaic genomes have been combined with one of the parental genotypes. In Figure 1, isolates VI354, K124, MAL, G141, and LBV10-5 and CI32 in the gag tree and isolates UG266, MAL, and K124 in the env tree are examples of this effect. Genotyping of such isolates requires further analysis.

Identification of mosaic genomes

Intergenotypic recombinants of HIV-1 consist of interspersed segments of genetic material from two or more parental genotypes. We have used a "boot scanning" procedure [8] whereby the available sequence is broken into overlapping segments 150–300 nt in length. Trees are then constructed from each segment using bootstrap resampling. Bootstrap values of 70% or greater provide reasonable confidence for assignment of an individual segment to one or the other genotype. Application of this method to the chimeric gag or env genes of isolates UG266, BZ200, respectively, is shown in Figure 2. Segments of the genes from these isolates cluster alternately with two different genotypes.

A more precise estimation of recombination breakpoints can subsequently be determined as described [4], using nucleotide sequence positions that differ between the two parental genotypes. Typically, breakpoints can be identified +/– 0.1 KB. Some segments of recombinant viruses cannot be assigned to either genotype with confidence (Figure 1, gray segments). These may represent regions of low information content or they could be regions with multiple crossovers. Conceivably, a currently undescribed genotype may have contributed to some mosaics.

Subgenomic sequence segments and genotypic assignments

The occurrence of mosaic or recombinant viral genomes complicates the derivation of genotypic assignments based on subgenomic sequence segments. It remains impractical to obtain full length genomic sequence of HIV-1 isolates as a routine genotyping method, due to the low abundance of HIV-1 proviral DNA in clinical samples and virus cultures on PBMC substrate, and to the relative inefficiency of the polymerase chain reaction when amplicons become large (however, see [9,10]). The great majority of genotypic assignments for HIV-1 are based on subgenomic sequence segments,

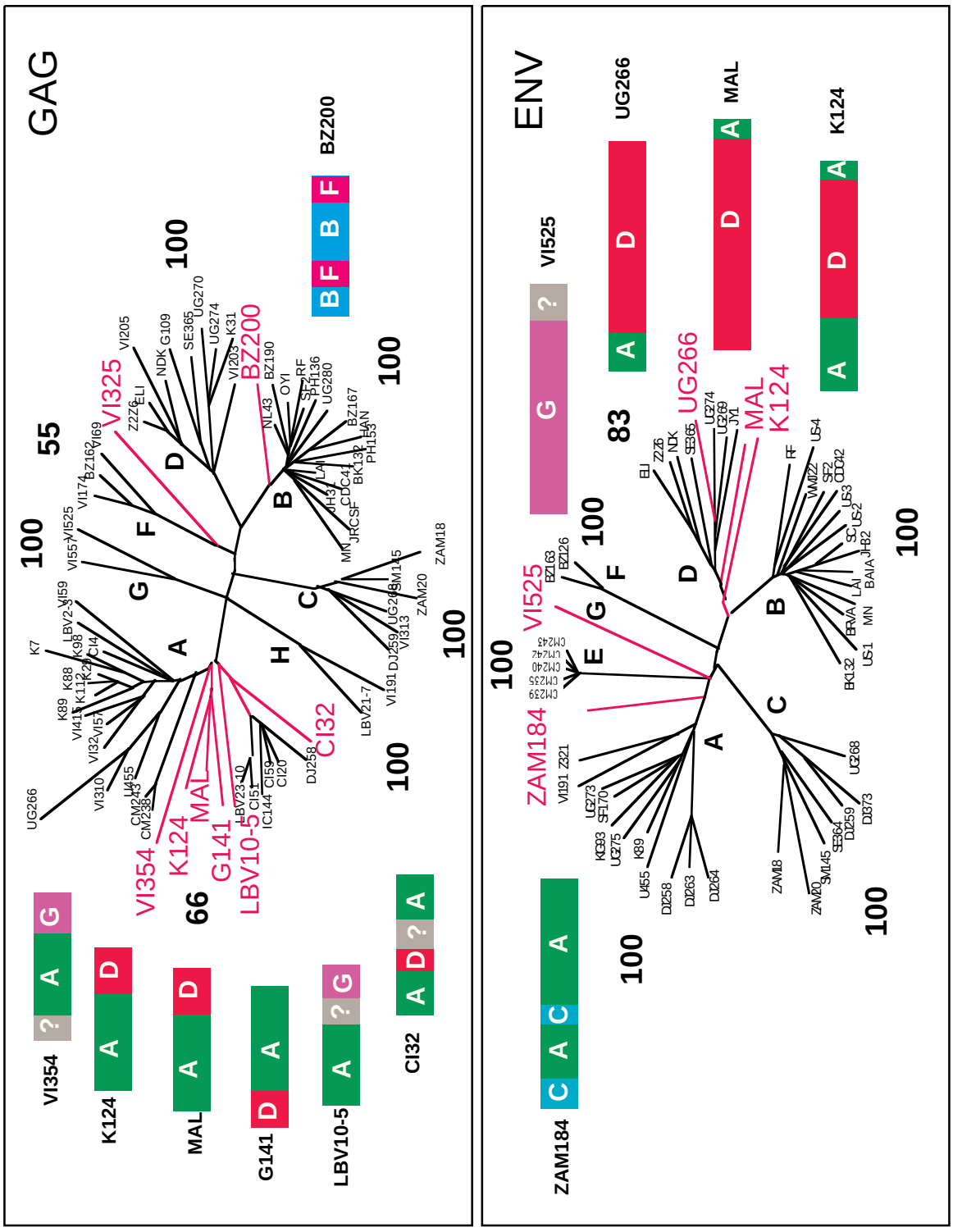

Figure 1. Phylogenetic trees and Mosaic Genomes. Neighbor-joining trees of HIV-1 gag and env genes [6,7] are shown. HIV-1 genotypes A through H and the bootstrap resampling frequencies for the main branches of the trees are indicated. Isolates in red are intergenic recombinants, whose structures are indicated in the accompanying diagrams.

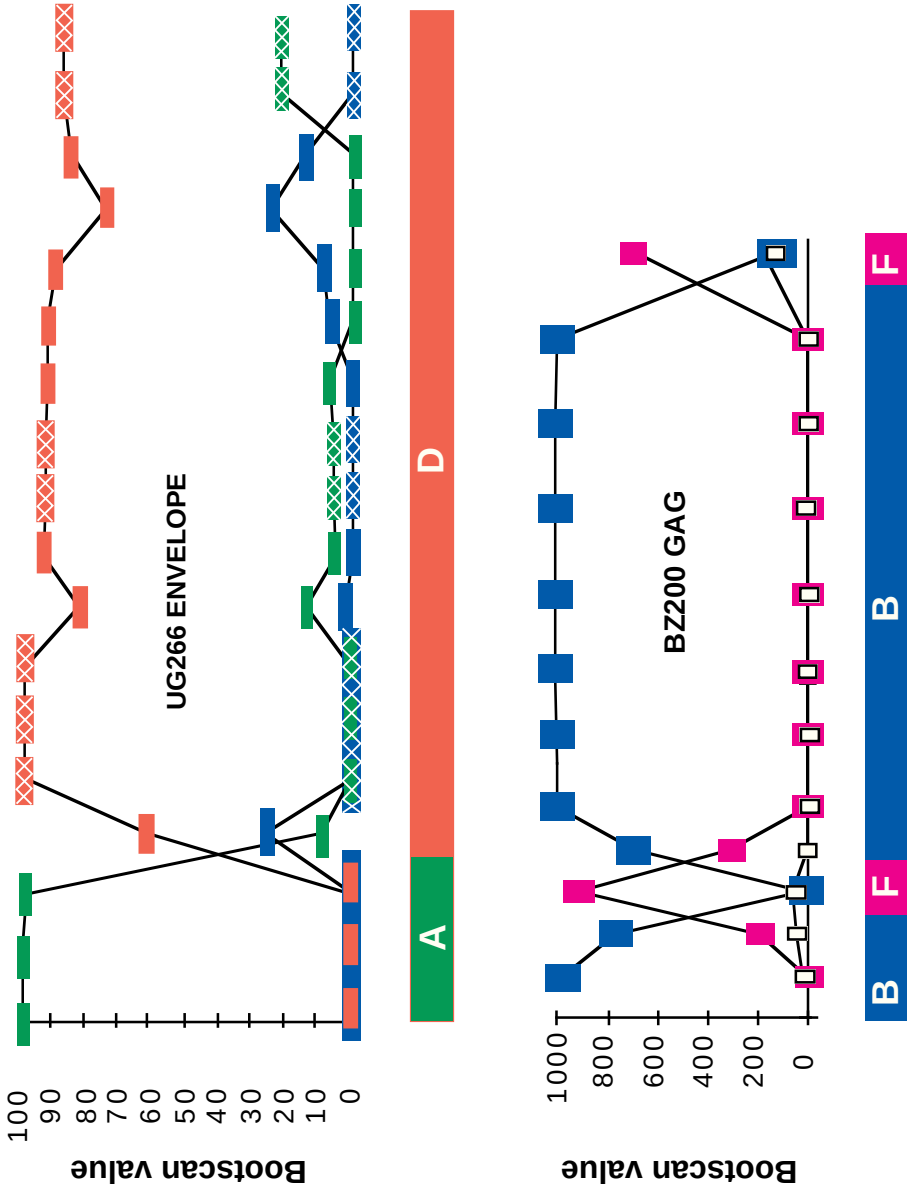

Figure 2. Bootscanning Procedure for Identification of Mosaic Genomes. DNA sequences of mosaic and parental genomes, together with an outgroup, have been aligned and used to build phylogenetic trees for sequential segments of 0.3 KB, each overlapping the previous by 0.15 KB. Essentially identical results were obtained using parsimony, maximum likelihood, and distance matrix-based algorithms. The upper panel shows bootstrap frequencies for 100 iterations using parsimony for the gene sequence of the full envelope of Ugandan isolate UG266. The lower panel shows a similar analysis of the gag gene of isolate BZ200 from Brazil. Here analysis was with maximum likelihood using 1000 iterations. In the upper panel, some initially unclassifiable segments were analyzed using larger segments (patterned boxes); the parental genotypes were A (green) and D (orange) with genotype B (blue) as the outgroup. The lower panel employed a uniform segment size and included genotypes B (blue) and F (magenta) with genotype H (white) as outgroup.

typically encompassing 2% to 30% of the genome. What fraction of the HIV-1 genome should be sequenced to provide sufficient data for genotyping?

At the present time there is no definitive answer to this question. It depends on the frequency with which breakpoints occur in mosaic genomes and on their distribution. Based on three full-length genomic sequences and on several additional isolates sequenced in gag and/or envelope genes, we have observed that the breakpoints occurred, on average, approximately once per kilobase. Insufficient data is available to determine whether breakpoints occur at a constant frequency across the genome, but, again, our unpublished data suggests that a multiplicity of breakpoints occur and that some locales are preferred. Obviously, the length of the sequence segment examined is a critical element of any genotyping effort, and longer sequence segments are preferable for the reasons outlined above. Multiple, short sequence segments do not necessarily provide the same power to resolve mosaic genomes as a single long segment, as each segment used in the bootscanning procedure must be of sufficient length to contain the information to distinguish genotypes.

"New" HIV-1 Subtypes

The multiple genotypes of HIV-1 were discovered sequentially over several years, and it is reasonable to anticipate the discovery of additional genotypes in the future. These are often termed "new", meaning that they have been discovered at a later time than "established" genotypes. In light of the occurrence of recombinant genomes, phylogenetic trees of HIV-1 isolates are sometimes not sufficient in themselves to distinguish new genotypes from intergenotypic recombinants. Mosaic sequences with multiple internal breakpoints seem to be the most problematic, as they can separate cleanly from established subtypes, leading to their mis-identification as new subtypes. The interpretation of sequences with a single breakpoint seems to be influenced by the proportional contribution of the parental genotypes. When only a small segment comes from the second genotype, a branch near the main trunk is usually established, whereas a more equal contribution from the two sources can position mosaics well within a genotype. Lastly, by inadvertently combining several mosaic genomes with the parental genotypes in a tree (for example, the A/D recombinants in the gag tree of Figure 1), the chimeras can become intertwined with the parental genotypes, leading to the impression of a single genotype with increased inter-isolate distances.

How should new genotypes be identified? Based on the experience to date, reasonable criteria might include: a) the appearance of at least two epidemiologically unrelated isolates that cluster together and that are separated from established genotypes; b) the availability of a least 1.5 kilobase of contiguous sequence from each; and c) the absence of any subsegment that can join established genotypes.

Summary

HIV-1 evolves both by the rapid accumulation of point mutations and by recombination [11–13]. At present, the relative contributions of these processes to the complex array of HIV-1 genotypes and to their phylogenetic relationships is unclear. Efforts to classify and organize HIV-1 genetic information may benefit from a heightened awareness that both processes are actively contributing to the diversity of viral forms.

References

[1] Myers G, Korber B, Wain-Hobson S, Jeang KT, Henderson LE, and Pavlakis GN. *Human Retroviruses and AIDS*, Los Alamos National Laboratory, Los Alamos, NM, 1994.

[2] Hahn BH, Robertson DL, McCutchan FE, Sharp PM. Recombination and Diversity of HIV: Implications for Vaccine Development. *Neuvième Colloque des Cent Gardes*, Paris, 1994.

[3] Sabino EC, Shaper EG, Morgado MG, Korber BTM, Diaz RS, Bongertz V, Cavalcante S, Galvao-Castro B, Mullins JI, Mayer A. Identification of Human Immunodeficiency Virus Type 1 Envelope Genes Recombinant Between Subtypes B and F in Two Epidemiologically Linked Individuals from Brazil. *J. Virol*. **68**: 6340–6346, 1994.

[4] Roberston DL, Sharp PM, McCutchan FE, Hahn BH. Recombination in HIV-1. *Nature* **374**: 124–126, 1995.

[5] Leitner T, Escanilla D, Marquina S, Wahlberg J, Brostrom C, Hansson HB, Uhlen M, Albert J. Biological and Molecular Characterization of Subtype D, G, and A/D Recombinant HIV-1 Transmissions in Sweden. *Virology* **209**: 136–146, 1995.

[6] Louwagie J, McCutchan F, Peeters M, Brennan TP, Sanders-Buell E, Eddy G, van der Groen G, Fransen K, Gershy-Damet GM, Deleys R, Burke DS. Phylogenetic Analysis of Gag Genes from 70 International HIV-1 Isolates Provides Evidence for Multiple Genotypes. *AIDS* **7**: 769–780, 1993.

[7] Louwagie J, Janssens W, Mascola J, Heyndricks L, Hegerich P, van der Groen G, McCutchan FE, Burke DS. Genetic Diversity of the Envelope Glycoprotein from Human Immunodeficiency Virus Type-1 (HIV-1) Isolates of African Origin. *J. Virol*. **69**:263–271, 1995.

[8] Salminen M, Carr JK, Burke DS, and McCutchan FE. Identification of Recombination Breakpoints in HIV-1 by Bootscanning. Laboratory of Tumor Cell Biology National Cancer Institute Meeting, Bethesda, MD, 1995.

[9] Salminen M, Carr J, Burke DS, McCutchan FE. Amplification and Cloning of Virtually Full-length HIV-1 Genomes from Diverse Subtypes. Keystone Symposium on HIV Pathogenesis, *J. Cell. Biochem. Suppl. 21B*, p. 247, 1995.

[10] Salminen MO, Koch K, Sanders-Buell E, Ehrenberg PK, Michael NL, Carr JK, Burke DS, McCutchan FE. Recovery of Virtually Full Length HIV-1 Provirus of Diverse Subtypes from Primary Virus Cultures Using the Polymerase Chain Reaction, *Virology* 1995, in press.

[11] Li WH, Tanimura M, Sharp PM. Rates and Dates of Divergence Between AIDS Virus Nucleotide Sequences. *Mol. Biol. Evol*. **5**:313–330, 1988.

[12] Hu WS, Temin HM. Retroviral Recombination and Reverse Transcription. *Science* **250**: 1227–1233, 1990.

[13] Coffin JM. Retroviridae and Their Replication. in: *Virology* (Fields BN and Knipe DM, eds.). Raven Press, pp. 1437–1500, 1990.